FINDING NONOVERLAPPING SUBSTRUCTURES OF A SPARSE MATRIX[*]

ALI PINAR[†] AND VIRGINIA VASSILEVSKA[‡]

Abstract. Many applications of scientific computing rely on computations on sparse matrices, thus the design of efficient implementations of sparse matrix kernels is crucial for the overall efficiency of these applications. Due to the high compute-to-memory ratio and irregular memory access patterns, the performance of sparse matrix kernels is often far away from the peak performance on a modern processor. Alternative data structures have been proposed, which split the original matrix A into A_d and A_s, so that A_d contains all dense blocks of a specified size in the matrix, and A_s contains the remaining entries. This enables the use of dense matrix kernels on the entries of A_d producing better memory performance. In this work, we study the problem of finding a maximum number of nonoverlapping rectangular dense blocks in a sparse matrix, which has not been studied in the sparse matrix community. We show that the maximum nonoverlapping dense blocks problem is NP-complete by using a reduction from the maximum independent set problem on cubic planar graphs. We also propose a 2/3-approximation algorithm for 2×2 blocks that runs in linear time in the number of nonzeros in the matrix. We discuss alternatives to rectangular blocks such as diagonal blocks and cross blocks and present complexity analysis and approximation algorithms.

Key words. Memory performance, memory-efficient data structures, high-performance computing, sparse matrices, independent sets, NP-completeness, approximation algorithms.

1. Introduction.

Sparse matrices lie in the hearts of many computation-intensive applications such as finite-element simulations, decision support systems in management science, power systems analysis, circuit simulations, and information retrieval. The performances of these applications directly rely on the performances of the employed sparse matrix kernels . The memory performance of sparse matrix operations on modern processors however, is often a bottleneck due to the irregular memory-access patterns of sparse matrix operations, and extra memory load operations required to exploit sparsity. The memory-performance bottleneck is becoming more crucial everyday, arguably becoming the most important problem in high performance computing. To overcome this memory bottleneck designing alternative data structures for sparse matrices that are memory friendly has been investigated. One common approach in these efforts is to exploit the special substructures in a sparse matrix, such as small dense matrices, to decrease the number of extra load operations. In this paper, we study the problem of finding a maximum number of nonoverlapping substructures in a sparse matrix, with the objective of improving the effectiveness of alternative sparse matrix data structures that exploit dense blocks.

Conventional data structures for sparse matrices have two components: an array that stores floating-point entries of the matrix and arrays that store the nonzero structure (i.e., pointers to the locations of the numerical entries). Exploiting sparsity invariably requires using pointers, but pointers often lead to poor memory performance. One reason for the poor memory performance is that pointers cause an irregular memory access pattern and thus poor spatial locality. Another important reason, which is often overlooked, is the extra load operations. Each operation on a nonzero entry requires loading the location of that nonzero before loading the actual

[*]This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract DE-AC03-76SF00098.

[†]Corresponding author. Computational Research Division, Lawrence Berkeley National Laboratory, (apinar@lbl.gov).

[‡] Computer Science Department, Carnegie Mellon University (virgi@cs.cmu.edu).

$$\begin{pmatrix} x & x & x & & \\ & x & & x & \\ x & x & & & \\ & & & x & x \\ & & x & x & x \end{pmatrix} = \begin{pmatrix} x & x & & & \\ & & & & \\ x & x & & & \\ & & & x & x \\ & & x & x & \end{pmatrix} + \begin{pmatrix} & & x & & \\ & x & & x & \\ & & & & \\ & & & & \\ & & & & x \end{pmatrix}$$

$$A \quad = \quad A_{12} \quad + \quad A_{11}$$

FIG. 1.1. *Matrix splitting.*

floating point number. For instance, sparse matrix vector multiplication, which is one of the most important kernels in numerical algorithms, requires three load operations for each multiply-and-add operation. And it has been observed that this overhead might be as costly as the floating point operations [5].

Recent studies have investigated improving memory performance of sparse matrix operations by reducing the number of extra load operations [5, 8, 9, 10]. Toledo [9] studied splitting the matrix as $A = A_{12} + A_{11}$, where A_{12} includes 1×2 blocks of the matrix (two nonzeros in consecutive positions on the same row), and A_{11} covers the remaining nonzeros, as illustrated in Fig. 1.1. Notice that it is sufficient to store a pointer for each block in A_{12}. In [8], Pınar and Heath studied the reordering problem to increase the sizes of these blocks. They proposed a graph model to reduce the matrix ordering problem to the traveling salesperson problem. Vuduc et al. studied various blocking techniques to decrease load operations, and improve cache utilization [10]. Significant speedups in large experimental sets have been observed, which gives motivation to search for larger blocks in the matrix for further improvements in performance. Splitting operation can be generalized to exploit arbitrary substructures. For instance, one can split the matrix into $A = A_d + A_s$, where A_d contains all specified substructures, and A_s contains the remaining entries. Clearly, for a specified substructure, having more entries in A_d merits fewer load operations, thus better memory performance. This calls for efficient algorithms to find a maximum number of nonoverlapping substructures in a sparse matrix. A greedy algorithm is sufficient to find a maximum number of nonoverlapping $m \times n$ dense matrices when $m = 1$ or $n = 1$. However, this problem is much harder when $m, n \geq 2$.

In this work, we study the problem of finding a maximum number of nonoverlapping substructures of a sparse matrix, which we call the *maximum nonoverlapping substructures* problem. We focus on $m \times n$ dense blocks as a substructure, due to their availability in sparse matrices arising in various applications, and effectiveness in decreasing extra load operations. We call this problem the *maximum nonoverlapping dense blocks* problem. In the next section, we define the problem formally and investigate its relation to the maximum independent set problem. We define a class of graphs where the independent set problem is equivalent to the maximum nonoverlapping dense blocks problem. In Section 3, we use this relation to prove that the maximum nonoverlapping dense blocks problem is NP-complete. Our proof uses a reduction from the maximum independent set problem on cubic planar graphs and adopts orthogonal drawings of planar graphs. Section 4 presents an approximation algorithm for the problem. Since we are motivated by improving memory performance of sparse matrix operations, we are interested in fast and effective heuristics for the preprocessing cost to be amortized over the speedups in subsequent sparse matrix operations. Our algorithms require only linear time and space, and generate solutions whose sizes are within 2/3 of the optimal. In Section 5, we discus alternative patterns to rectangular blocks. We show that the problem of finding diagonal blocks can be reduced to that of finding rectangular blocks, and thus the problem is NP-complete, and our 2/3-approximation algorithm is valid for diagonal blocks as well. We also discuss the cross blocks, prove that finding a maximum set of nonoverlapping cross blocks is NP-complete, and generalize our results for variations of the cross block. We present some open problems in Section 6 and conclude with Section 7.

The problem of finding nonoverlapping dense blocks of a sparse matrix has not been studied in the sparse-matrix community. We have been recently aware of the work by Berman *et al.* [2], where a similar problem is discussed as the optimal tile salvage problem. In the optimal tile salvage problem, we are given an $\sqrt{N} \times \sqrt{N}$ region of the plane tiled with unit squares, some of which have been removed. The task is to find a maximum number of functional nonoverlapping $m \times n$ tiled rectangles. The difference between our problem and the optimal tile salvage problem is that in the tile salvage problem the tiles are allowed to be in any orientation ($m \times n$ or $n \times m$), whereas in our case the orientation is fixed (only $m \times n$). The two problems coincide in the case of square dense blocks. Berman *et al.* proved the NP-completeness of the tile salvage problem, however their proof exploits the flexibility in the orientation of the dense block, and thus our proof is significantly different. Berman *et al.* also describe an $(1 - \epsilon)$-approximation algorithm, which would work for square blocks, for $\epsilon = O(1/\sqrt{\delta \log M})$, where M is the optimal solution value. Their algorithm is based on maximum planar H-matching which runs in $O(N^{1+\delta})$ steps for small $\delta > 0$. Baker [1] also has an algorithm for the case of square blocks, which runs in $O(8^k N)$-time and $O(4^k N)$ space and produces a $(k-1)/k$-approximation. Both of these algorithms however are complex and hard to implement. The greedy 2/3-approximation algorithms we propose are very simple. It requires linear time and space, with very small constant factors in the time and space bounds. Our algorithm requires only one pass through the matrix, and thus is I/O-efficient.

2. Preliminaries.

In this section we define the problems formally, and present definitions and some preliminary results that will be used in the following sections.

2.1. Problem Definition.

This work investigates the problem of finding a maximum number of nonoverlapping matrix substructures of prescribed form and orientation.

Definition 2.1. *An $m \times n$* pattern *is a 0-1 $m \times n$ matrix σ. An* oriented σ-substructure *of a matrix A is an $m \times n$ submatrix M in A so that $M(i,j) \neq 0$ if $\sigma(i,j) = 1$ for $1 \leq i \leq m$, and $1 \leq j \leq n$. Two substructures M and N overlap if they share nonzero entry e in M with coordinates (i_M, j_M) in M and (i_N, j_N) in N and $\sigma(i_M, j_M) = \sigma(i_N, j_N) = 1$.*

Given a particular pattern σ, we define the *maximum nonoverlapping σ-substructures* (MNS) problem as follows.

> *Given an $M \times N$ matrix A and integer K, does A contain K disjoint σ-substructures?*

In this paper, we mostly focus on *dense blocks*, due to their simplicity, and their effectiveness in speeding up sparse matrix operations. A *dense block* of a matrix is a submatrix of specified size all of whose entries are nonzero, i.e., it is a σ-substructure where σ is the all 1s matrix. We identify a dense block with its upper left corner. Two blocks overlap if they share a matrix entry. Formally, *Given an $M \times N$ matrix $A = (a_{ij})$, we say b_{ij} is an $m \times n$ dense block in A iff $a_{kl} \neq 0$ for all k and l such that $i \leq k < i+m \leq M$ and $j \leq l < j+n \leq N$. Two $m \times n$ blocks b_{ij} and b_{kl} overlap iff $i \leq k < i+m$ and $j \leq l < j+n$, or $k \leq i < k+m$ and $l \leq j < l+n$.*

We define the *maximum nonoverlapping dense blocks* (MNDB) problem, which restricts the MNS problem to dense blocks as follows.

Given an $M \times N$ matrix A, positive integers m and n that define the block size, and a positive integer K, does A contain K disjoint $m \times n$ dense blocks?

2.2. Intersection Graphs. It is easy to find all specified patterns in a matrix, however what we need is a subset with nonoverlapping blocks. In this sense, the MNS problem is related to the *maximum independent set* (MIS) problem, which is defined as finding a maximum cardinality subset of vertices I of a graph G, such that no two vertices in I are adjacent. Below we define an *intersection graph*, which reveals the relation between the independent set and the nonoverlapping blocks problems more clearly.

DEFINITION 2.2. *A graph G is an intersection graph of the σ-substructures of a matrix A if there is a bijection ϕ between the vertices of G and the substructures of A, such that there is an edge in G between $\phi(s_1)$ and $\phi(s_2)$ if and only if s_1 and s_2 overlap in A.*

We will use $G(A, m, n)$ to refer to the intersection graph of dense $m \times n$ blocks in matrix A. A maximum independent set on $G(A, m, n)$ gives a maximum number of nonoverlapping blocks in A, thus the MNDB problem can be reduced to the maximum independent set problem, which is known to be NP-complete [4]. However it is important to note that the block intersection graphs have special structures, which can be exploited for efficient solutions. For instance, a greedy algorithm is sufficient to find a maximum number of nonoverlapping $1 \times n$ and $m \times 1$ blocks, since these problems reduce to a family of disjoint maximum independent set problems on interval graphs. In the remainder of this section, we define the class of graphs that constitute block intersection graphs. An intersection graph of a set of 2×2 dense blocks is an induced subgraph of the so called X-grid which consists of the usual 2 dimensional grid, and diagonals for each grid square. In general, the intersection graph of a set of $m \times n$ dense blocks is an induced subgraph of the X_{mn} grid. Below, we first define an X_{mn} grid, and then restrict the definition to define the graph class $X\Gamma_{mn}$ that represent graphs that can be an intersection graph for a matrix.

DEFINITION 2.3. *An $M \times N$ X_{mn} grid is a graph with a vertex set V and an edge set E, so that*

- $V = \{v_{ij} : 1 \leq i \leq M - m; 1 \leq j \leq N - n + 1\}$
- $E = \{(v_{ij}, v_{kl}) : 1 \leq i, k \leq M - m + 1; 1 \leq j, l \leq N - n + 1 : |i - k| < m; |j - l| < n\}$

In an X_{mn} grid, vertex v_{ij} corresponds to the block b_{ij} in the matrix, and edges correspond to all possible overlaps between blocks. Note that not all induced subgraphs of the X_{mn} grid are intersection graphs of a matrix. We define a graph class $X\Gamma_{mn}$ in which each graph corresponds to an intersection graph $G(A, m, n)$ of the set of $m \times n$ dense blocks of a matrix A, and each such intersection graph is in the class.

Definition 2.4. *A graph $G = (V, E)$ is in the graph class $X\Gamma_{mn}$ if and only if it is an induced subgraph of an X_{mn} grid and has the closure property so that $v_{ij} \in V$ if*

$$\forall i \le k < i+m, j \le l < j+n; \quad \exists v_{st} : s \le k < s+m \text{ and } t \le l < t+n$$

The closure property enforces that there is a vertex in the graph for each block in the matrix. Being an induced subgraph of an X grid guarantees that there is an edge for each overlap. The graphs in this class are exactly the intersection graphs of the $m \times n$ blocks in a matrix, thus finding a maximum independent set of a graph in this class is equivalent to solving the MNDB problem of the corresponding dense matrix blocks. This claim is formalized by the following lemma.

Lemma 2.1. *An instance of the MNDB problem for finding $m \times n$ nonoverlapping dense blocks in a matrix A is polynomially equivalent to an instance of MIS for a graph in $X\Gamma_{mn}$.*

Proof. As we discussed earlier, the MNDB problem can be reduced to the problem of finding an independent set on its intersection graph. Here we show the one-to-one correspondence between intersection graphs, and graphs in $X\Gamma_{mn}$. Remember that each dense block b_{ij} corresponds to the vertex v_{ij} in $G(A, m, n)$. By definition of the class $X\Gamma_{mn}$, $G(A, m, n) \in X\Gamma_{mn}$, thus any instance of an MNDB problem can be reduced to an independent set problem in a graph in $X\Gamma_{mn}$.

Given a graph G in $X\Gamma_{mn}$, define $A = (a_{ij})$, so that a_{ij} is a nonzero iff $k \le i < k+m$ and $l \le j < l+n$ for some vertex v_{kl} in G. Observe that any dense block in A must be represented by a vertex in G due to the closure property. Also, for any two adjacent vertices in G, corresponding blocks intersect in A, and no other blocks overlap, due to the definition of edges in X_{mn}. Thus, a maximum-cardinality subset of nonoverlapping blocks in matrix A corresponds to a maximum independent set in $G \in X\Gamma_{mn}$. $\square$

In this paper we will use the graph class $X\Gamma_{22}$ to prove the NP-completeness of the MNDB problem for 2×2 blocks. Our proof can be generalized to arbitrary sized blocks, showing the NP-completeness of the MNDB problem for $m \times n$ blocks, and hence the NP-completeness of the maximum independent set problem for graphs in class $X\Gamma_{mn}$.

The following lemma shows that removing a subset of the vertices along with their neighbors preserves the characteristics of the graph, providing the basis for greedy approximation algorithms as will be presented in Section 4.

Lemma 2.2. *Let $G = (V, E)$ be a graph in $X\Gamma_{mn}$, $S \subseteq V$ a subset of vertices, and $N(S) = \{u \mid (u, v) \in E,\ v \in S,\ u \notin S\}$ be the neighborhood of S in G. Then the graph G' induced by $V \setminus (S \cup N(S))$ is still in $X\Gamma_{mn}$.*

Proof. Removing a vertex and its neighbors in G corresponds to removing all nonzeros in a bl ock in the corresponding matrix. The remaining graph is the intersection graph of the resulting matrix. $\square$

2.3. Planar Graphs and Orthogonal Drawings. A graph G is *planar* if and only if there exists an embedding of G on the sphere such that no two edges have a point in common besides the vertices. G is *cubic planar* if every vertex has degree 3.

An *orthogonal drawing* of a graph G is an embedding of G onto a 2-dimensional rectangular grid such that every vertex is mapped to a grid point and every edge is mapped to a continuous path of grid line segments connecting the end points of the

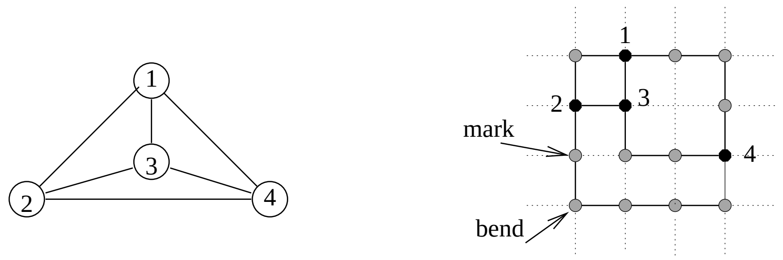

FIG. 2.1. *Planar orthogonal drawing*

edge. When G is planar, the edge paths do not cross. An example of orthogonal embedding of a planar graph is illustrated in Fig. 2.1. As seen in this figure, we refer to a grid point where an edge path changes direction as a *bend.* No two edges share a grid segment or a bend, and no edge path can go through a vertex unless this vertex is an end point of the edge corresponding to the path and is an end point of the path itself. A *mark* in an orthogonal drawing of a graph is a grid point that an edge passes through,but not a vertex in the original graph. The following result has been reported by de Fraysseix *et al.* [3], Kant [6], and Papakostas and Tollis [7].

THEOREM 2.3. *Every planar graph G with vertex degree at most 4 can be drawn orthogonally with at most $\lfloor \frac{n}{2} \rfloor + 1$ bends on an $\lfloor \frac{n}{2} \rfloor \times \lfloor \frac{n}{2} \rfloor$ grid in linear time.*

In particular, this shows that every cubic planar graph $G = (V, E)$ can be embedded orthogonally in an $O(|V|) \times O(|V|)$ grid in polynomial time. The NP-completeness proof in the next section uses a reduction from the maximum independent set (MIS) problem on cubic planar graphs, and adopts orthogonal drawings.

3. Complexity.

This section proves that the MNDB problem is NP-complete for 2×2 blocks. We use a reduction from the independent set problem on cubic planar graphs, which is NP-complete [4]. Throughout this section, we let $X\Gamma$ denote $X\Gamma_{22}$. The next lemma explains how we can retain independent set characteristics of the problems after transformations.

LEMMA 3.1. *Let $G = (V, E)$ be a graph, and u, v be two adjacent vertices in G, so that all neighbors of u besides v are also neighbors of v. Let $G' = (V', E')$ be the graph G after vertex v is removed. The size of the maximum independent set in G is equal to the size of the maximum independent set in G'.*

Proof. If vertex v is in a maximum independent set I, then none of its neighbors are in I. Thus $I' = I \cup \{v\} \setminus \{u\}$ is an independent set in G and in G' of the same size as I. $\square$

COROLLARY 3.2. *Let $G \in X\Gamma$ contain the graph H in Fig. 3.4(a) as an induced subgraph so that all vertices except for possibly v_1, v_2 and v_3 have all of their neighbors in H. Then any instance (G, K) of MIS is equivalent to the instance (G', K) of MIS for the graph $G' = G \setminus \{w_1, w_2\}$.*

Proof. By Lemma 3.1, we can remove w_1 from the graph since all neighbors of x_1 are neighbors of w_1 as well. The reduced graph is illustrated in Fig. 3.4(b). Again using Lemma 3.1, we can remove w_2 since it covers all neighbors of x_2. Note that we can apply the same transformation to add vertices w_1 and w_2 to the graph in Fig. 3.4(c). $\square$

The following lemma describes how edges of a graph can be replaced by paths, while preserving independent set characteristics.

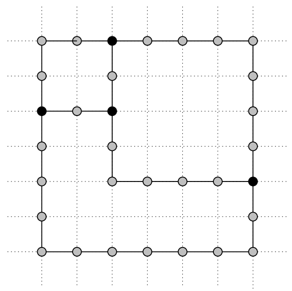

FIG. 3.1. *Enlargement operation for $K = 1$*

LEMMA 3.3. *Let $G = (V, E)$ be a graph and $e = (v_i, v_j) \in E$ be an edge. Let $G_{e,k}$ be the graph G with the edge e substituted by a simple path $v_i, w_1, w_2, \ldots, w_{2k}, v_j$ where $k \in \mathbf{Z}^+$ and w_i are new vertices not in the original graph. Then there exists an independent set of size K in G if and only if there exists an independent set of size $K + k$ in $G_{e,k}$.*

Proof. We present the proof for $k = 1$, and the result follows by induction.
Sufficiency: Let I be an independent set in G, then either $v_i \notin I$ or $v_j \notin I$. Without loss generality, assume $v_i \notin I$, then $I' = I \cup \{w_1\}$ is an independent set in $G_{e,k}$.
Necessity: Let I' be an independent set in $G_{e,k}$. If $w_1 \in I'$, then $v_i \notin I'$, thus $I = I' \setminus \{w_1\}$ is an independent set in G. Symmetrically, if $w_2 \in I'$, then $v_j \notin I'$, thus $I = I' \setminus \{w_2\}$ is an independent set in G. If $w_1, w_2 \notin I'$, then $I = I' \setminus \{v_2\}$ is an independent set in G. $\square$

THEOREM 3.4. *Problem MNDB is NP-complete for* 2×2 *blocks.*

Proof. As discussed in the previous section, the problem of finding maximum number of nonoverlapping dense blocks in a sparse matrix can be reduced to the problem of finding a maximum independent set in the intersection graph of the matrix, and thus is in NP. For the NP-completeness proof we use reduction from the independent set problem on cubic planar graphs, which is NP-complete [4]. We first use Theorem 2.3 to embed a cubic planar graph onto a grid. Then we transform the embedded graph so that it is in $X\Gamma$. Our transformations preserve independent set characteristics so that an independent set in the transformed graph can be translated to an independent set in the original graph. Finally, we use Lemma 2.1 to relate the independent set problem on a graph in $X\Gamma$, to the MNDB problem, and conclude the MNDB problem is NP-complete.

Our transformations are local, so we first *enlarge* the grid to make room for these transformations. The enlargement operation inserts K new grid points between two grid points in the original. An example is illustrated in Fig. 3 for $K = 1$. After the enlargement, each edge is now replaced by a path of K vertices (which we distinguish from the original vertices by calling them marks). Two adjacent vertices in the original graph are now at a distance $K + 1$, which generates a $K \times K$ area around each vertex for local transformations. In this proof, it is sufficient to use $K = 100$.

We can break down our transformations into 2 steps. The first step guarantees that the transformed graph is in $X\Gamma$. For this purpose, we need to have an edge between all pairs of vertices for which the corresponding blocks overlap so that the graph is in $X\Gamma$, and we need to insert vertices into the graph if necessary so that the closure property is satisfied. The second step makes sure that each edge in the original graph is replaced by an even-length path after the orthogonal embedding and

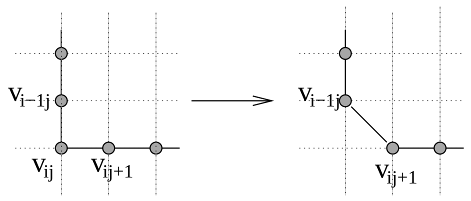

FIG. 3.2. *Bend transformation*

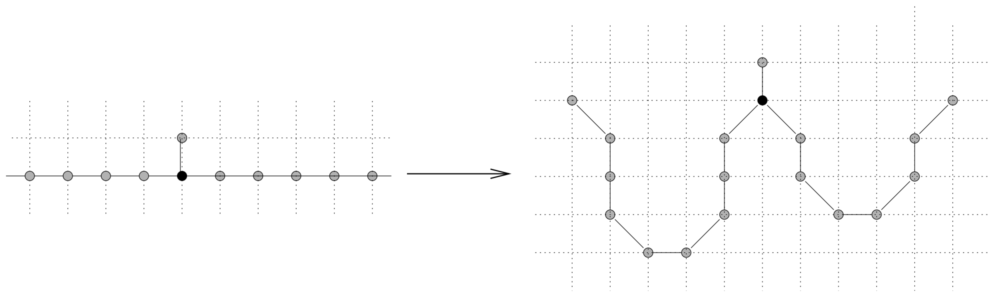

FIG. 3.3. *T-junction transformation*

transformations. Then we have successfully transformed the independent set problem on the cubic planar graph to an independent set problem on a graph in $X\Gamma$, and we can conclude the NP-completeness of the MNDB problem using the result of Lemma 2.1.

We need to consider two cases for the first step. One is a *bend* neighborhood as illustrated in Fig. 3.2, and the other is a *T- junction.* As illustrated in Fig. 3.3 a T-junction is just a neighborhood of a vertex in the original graph. Notice that the only remaining case is a path of vertices, which does not cause any problems. Consider a bend v_{ij} connected to two other marks v_{i-1j} and v_{ij+1}. Note that v_{ij} cannot be a vertex in the original graph, since the original graph is cubic. In a graph in $X\Gamma$, there must be and an edge between v_{i-1j} and v_{ij+1}. We can remove v_{ij}, and connect v_{i-1j} and v_{ij+1} as in Fig. 3.2.

Now consider a T-junction with vertex v_{ij} at the center, as illustrated in Fig. 3.3. The neighborhood of v_{ij} is composed of (up to a rotation) v_{ij-1}, v_{ij+1}, and $v_{i-1\,j}$, none of which is a vertex in the original graph. As in the case of a bend, the problem here is the absence of edges between v_{ij-1} and $v_{i-1\,j}$, and between $v_{i-1\,j}$ and v_{ij+1}, for which the associated blocks will overlap. Also observe that v_{ij} must be a vertex of the original graph, and cannot be eliminated. We can make the transformation illustrated in Fig. 3.3, yet the resulting graph is still not in $X\Gamma$, since it has missing vertices, and does not satisfy the closure property. We can use Corollary 3.2 to add vertices to the graph as depicted in Fig. 3.4, so that the resulting graph is in $X\Gamma$.

By Lemma 3.3, we need each path replacing an edge of the planar graph to be of even length. For each edge going through an odd number of marks we know that there is a straight line segment going through at least 7 marks, due to the initial enlargement. We can replace this 7 vertex segment with an 8 vertex segment, to guarantee that the path representing an edge is of even length. This transformation is illustrated in Fig. 3.5. After this step, we have a graph in $X\Gamma$ that replaces each edge in the original graph with an even length path.

Notice that all our transformations require polynomial time and space, thus the

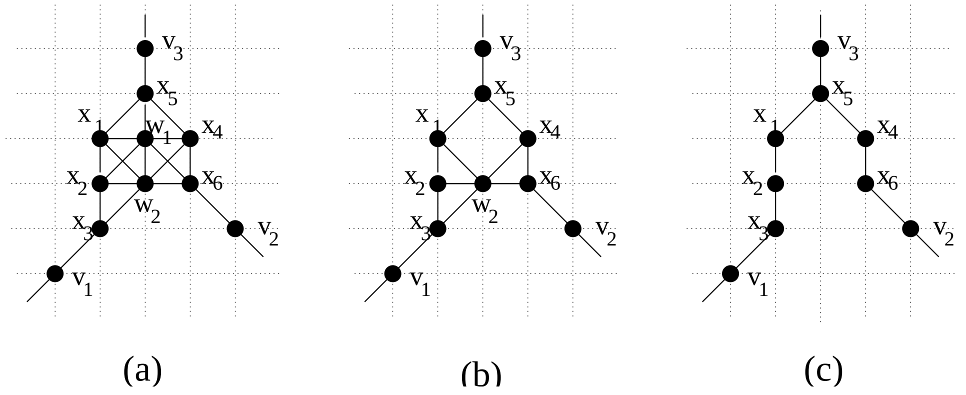

FIG. 3.4. *Transformation to preserve closure properties*

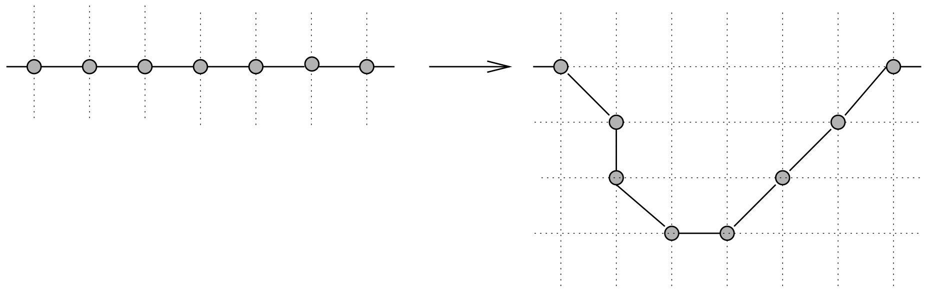

FIG. 3.5. *Odd-to-even length transformation to preserve independent set characteristics.*

size of the final embedded graph is polynomial in the size of the original graph.

This reduces the independent set problem for cubic planar graphs to an independent set problem in a graph in class $X\Gamma$. By the result of Lemma 2.1, we know the independent set problem on a graph in $X\Gamma$ is equivalent to a MNDB problem in a matrix. Thus we reduced the independent set problem for cubic planar graphs to the MNDB problem, which concludes our proof. ☐

Our proof serves as a template to prove NP-completeness of alternative substructures. Below, we generalize our result for arbitrary $m \times n$ blocks. In Section 5, we will use the same template to prove NP-completeness of the MNS problem for cross blocks.

THEOREM 3.5. *Problem MNDB is NP-complete for $m \times n$ blocks for $m, n > 2$.*

Proof. The reduction is again from the independent set problem on 3-planar graphs, and our proof uses only a minor modification to the proof of Theorem 3.4. Given a 3-planar graph G_P, we use exactly the same transformation as in Theorem 3.4, so that we have a graph $G = (V, E) \in X\Gamma_{22}$. What we need a is a graph in $X\Gamma_{mn}$. We will map vertices and edges G, which is on an $(M - m + 1) \times (N - n + 1)$ grid, onto an $[(M - m + 1)(m - 1) + 1] \times [(N - n + 1)(n - 1) + 1]$ X_{mn} grid to attain $G' = (V', E') \in X\Gamma_{mn}$. Our mapping stretches the graph so that overlaps of $m \times n$ blocks are minimal. That is blocks on the same row (column) overlap at $m \times 1$ $(1 \times n)$ blocks if they overlap. All other blocks overlap at 1×1 blocks at most. Each vertex in V' is an image of a vertex in V, so that $v_{ij} \in V$ is mapped to the vertex position $(i * (m - 1), j * (n - 1)$ in G'. Similarly, all edges in E' are images of edges in E, so that two vertices in V' are connected if and only if counterparts are connected in G.

The two graphs G and G' are essentially the same, thus an independent set on one can be trivially translated to an independent set on the other. Also $G' \in X\Gamma_{mn}$, since it contains edges for all potential overlaps. This concludes that the independent set problem on a 3-planar graph can be translated to an independent set problem on a graph in $X\Gamma_{mn}$, and thus the MNDB problem on a sparse matrix. $\square$

4. Approximation Algorithms.

In this section, we present a 2/3-approximation algorithm for the MNDB problem for 2×2 blocks. Now that we know the problem is NP-complete, we have to resort to heuristics for a fast and effective solution. Remember that our motivation for investigating this problem is speeding up sparse matrix-vector multiplication. Our methods will be used in a preprocessing phase, thus they must be fast, for their cost to be amortized by the speedup in subsequent sparse matrix-vector multiplications.

Berman *et al.* [2], propose an approximation algorithm for square blocks, which uses the Lipton-Tarjan planar separator algorithm to get a $(1-\epsilon)$-approximation, where $\epsilon = O(1/\sqrt{\delta log M})$ in $O(n^{1+\delta})$ time, for any $\delta > 0$, where M is the size of an optimal solution. Baker [1] gives an $(k-1)/k$-approximation, which uses $O(8^k n)$ time and $O(4^k n)$ space.

Below we propose a greedy approach for the 2×2 case, which in the 1/2-approximation case is applicable to general $m \times n$ rectangular blocks. Unlike the two algorithms cited, due to its greedy nature it is simple and very easy to implement. It is pass-efficient, and takes time and space linear in the number of blocks of the matrix, with very small constant factors in the bounds.

First note that an easy 1/2-approximation to the MNDB problem with 2×2, which runs in linear time in the number of blocks, is achieved by finding the leftmost block in the topmost row, adding it to the current independent set, and then repeating the same operation after removing this vertex and all its neighbors. Note that at most two of the vertices can be independent among those removed from the graph, thus we have a 1/2-approximation algorithm. In this section we show how to improve this approximation result by looking at an extended neighborhood of the leftmost vertex in the uppermost row. Our algorithm is based on choosing a set of vertices in the neighborhood of the leftmost vertex in the uppermost row, so that the size of this set is no less than 2/3 of a maximum independent set in the induced subgraph of those vertices removed from the graph. Clearly this generates a final solution that is 2/3 of the optimal, since all greedy decisions are at least 2/3 of the local optimal. Note that the resulting graph after removing a vertex along with all its neighbors still has the characteristics of the original as proven in Lemma 2.2

Our decision process **BinTreeDecision** is depicted as a binary decision tree in Fig. 4.1. In this tree, internal nodes indicate conditions, and the leaves list the vertices added to the independent set. We present the pseudocode of the algorithm below.

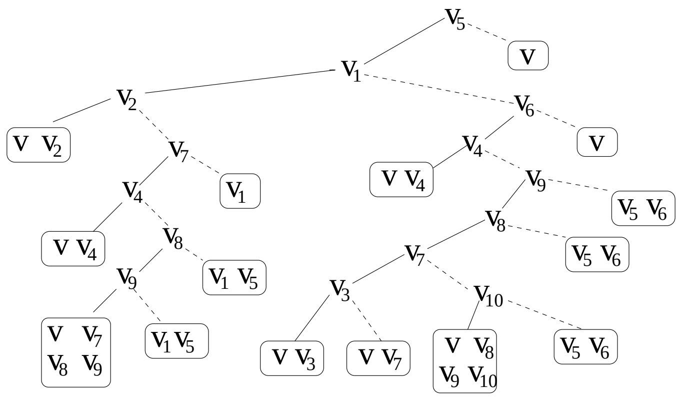

FIG. 4.1. *Decision tree for algorithm MNDB-APX. v corresponds to the leftmost vertex in the uppermost row, and the neighboring vertices in the X-grid are marked in Fig. 4.2. We take the left branch if the label vertex is in V, and the right branch otherwise. We proceed until we reach a leaf, which contains the set S that will be added in the independent set.*

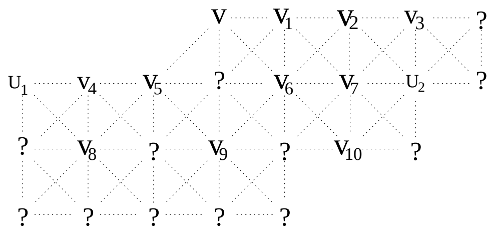

FIG. 4.2. *Vertex neighborhood considered for each call to BinTreeDecision. The positions v_i are used in the decision tree, while the positions u_i are only used in the analysis.*

```
Algorithm MNDB-APX
$I \leftarrow \emptyset$
  while $V \neq \emptyset$
    $v \leftarrow$ leftmost vertex on the uppermost row
    $S \leftarrow$ BinTreeDecision($v$)
    $I \leftarrow I \cup S$
    remove $S$ and its neighborhood from $G$
  endwhile
return $I$
```

LEMMA 4.1. *Algorithm MNDB-APX runs in linear time in the number of blocks in the matrix.*

Proof. Each iteration of the algorithm requires a traversal of the binary decision tree from the root to a leaf, which takes at most 8 steps, thus $O(1)$ time. Also at least one vertex is removed from the graph at each iteration. Thus the time for the decision process is linear in the number of vertices in the graph. The only other operation that affects the cost is finding the leftmost vertex in the uppermost row. In a preprocessing step one can go through the matrix in a left to right fashion and store pointers to the blocks so that v_{ij} appears before v_{kl} iff $i < k$ or $i = k$ and $j < l$. After this it takes constant time to find the current leftmost vertex on the uppermost row. ☐

Lemma 4.2. *The size of the maximal independent set returned by Algorithm MNDB-APX is no smaller than 2/3 of the size of maximum independent set on the intersection graph.*

Proof. The proof is based on case by case analysis. We show that **BinTreeDecision**(v) of Fig. 4.1 always returns an independent set S such that $N(S)$ contains no independent set larger than 1.5 $|S|$, where $N(S)$ denotes the neighborhood of S, i.e., the set of vertices in S or adjacent to a vertex in S. Below we examine the binary search tree case by case:

- $v_5 \notin V$ $S = \{v\}$, and v and its neighbors form a clique with MIS size 1.
- $v_5 \in V$
  - $v_1 \notin V$ By the closure property $v_2 \notin V$, and we have the following:
    - $v_6 \notin V$ $S = \{v\}$, and v and its neighbors form a clique with MIS size 1.
    - $v_6 \in V$
      - $v_4 \in V$ $S = \{v, v_4\}$, and $N(S)$ has MIS size at most 3.
      - $v_4 \notin V$ By the closure property $u_1 \notin V$. In this case, if one of v_9 or v_8 is not in V, then $S = \{v_5, v_6\}$, since their neighborhood has MIS size at most 3. Otherwise, $v_8, v_9 \in V$:
        - $v_7 \notin V$ This implies $u_2 \notin V$ and:
          - $v_{10} \notin V$ $S = \{v_5, v_6\}$ and $N(S)$ has MIS size at most 3.
          - $v_{10} \in V$ $S = \{v, v_8, v_9, v_{10}\}$, and $N(S)$ has MIS size at most 6.
        - $v_7 \in V$
          - $v_3 \in V$ $S = \{v, v_3\}$, and $N(S)$ has MIS size at most 3.
          - $v_3 \notin V$ $S = \{v, v_7\}$, and $N(S)$ has MIS size at most 3.
  - $v_1 \in V$
    - $v_2 \in V$ $S = \{v, v_2\}$, and $N(S)$ has MIS size at most 3.
    - $v_2 \notin V$ By the closure property $v_3 \notin V$, and
      - $v_7 \notin V$ $S = \{v_1\}$, v_1 and its neighbors form a clique, and the MIS is of size 1.
      - $v_7 \in V$
        - $v_4 \in V$ $S = \{v, v_4\}$, and $N(S)$ has MIS size at most 3.
        - $v_4 \notin V$ By the closure property $u_1 \notin V$, and if one of v_8 or v_9 is not in V, then $S = \{v_1, v_5\}$, and $N(S)$ has a MIS size at most 3. Otherwise if $v_8, v_9 \in V$, then $S = \{v, v_7, v_8, v_9\}$, and $N(S)$ has MIS size at most 6.

☐

Theorem 4.3. *Algorithm MNDB-APX is a linear time, 2/3-approximation algorithm for the MNDB problem.*

Proof. Follows directly from Lemma 4.1 and Lemma 4.2. ☐

Generalization of our 2/3-approximation algorithm for larger blocks is still under investigation. We expect the runtime and the approximation ratio to depend on the block size.

5. Alternative Substructures.

We have so far focused our discussions on finding dense rectangular blocks in a matrix. In this section, we will discuss generalizations of our results to alternative substructures that might be exploited to improve memory performance. We will first discuss diagonal blocks. Then we will introduce a *cross* substructure and its variants, and prove that MNS problem is NP-complete for finding these substructures.

5.1. Diagonal Blocks.

In many applications, nonzeros of the sparse matrix are lined around the main diagonal in the form of long diagonals. This makes diagonal

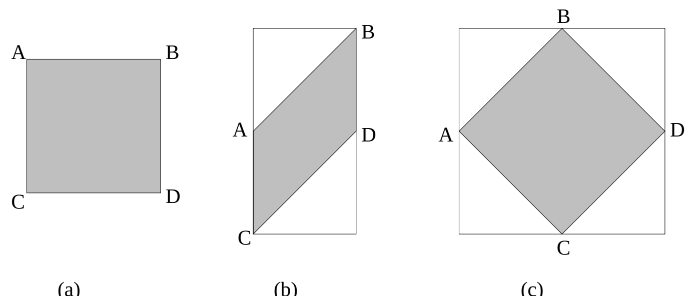

FIG. 5.1. *Matrix rotations. (a) the original matrix, (b) after Rotation 1, (c) after Rotation 2.*

blocks a nice alternative to rectangular blocks. We define a diagonal block as follows. Given an $M \times N$ matrix $A = (a(i,j))$, we say $d(i,j)$ is an $m \times n$ diagonal block in A iff

$$\forall k,l;\;\; i \le l < i+m;\; 0 \le k < n;\;\; a(l+k, j+k) \ne 0.$$

To find diagonal blocks in a sparse matrix, we can rotate the diagonals to transform diagonal blocks to rectangular blocks, so that our results for rectangular blocks can be applied directly. Our rotation is depicted in Fig. 5.1, and we define it as follows.

Rotation 1: *Given an $M \times N$ matrix A, its rotated matrix A_R is an $M+N-1 \times N$ matrix so that $A(i,j) \ne 0$ iff $A_R(i+N-j-1, j)$ for $0 \le i < M$ and $0 \le j < N$.*

THEOREM 5.1. *Given matrix A, let A_1 be its rotated matrix under Rotation 1. $d(i,j)$ is a diagonal block in A, iff $d(i+N-j-1, j)$ is a rectangular block in A_1.*

Proof. Necessity: Let $d(i,j)$ be a diagonal block in A. By definition of a diagonal block, and definition of Rotation 1, after transformation, we will have

$$\forall k,l;\;\; i \le l < i+m;\; 0 \le k < n;\;\; A_1(l+N-j-1, j+k) \ne 0$$

$$\Longrightarrow \forall k,l;\;\; 0 \le l < m;\; 0 \le k < n;\;\; A_1(i+N-j-1+l, j+k) \ne 0.$$

Thus $d(i+N-j-1)$ is an $m \times n$ rectangular block in A_1.

Sufficiency: Let $d(i+N-j-1, j)$ be an $m \times n$ rectangular block in A_1. This means before Rotation 1 we had,

$$\forall k,l;\;\; 0 \le l < m;\; 0 \le k < n;\;\; A(i+N-j-1-N+j+1+l+k, j+k) \ne 0$$

$$\Longrightarrow \forall k,l;\;\; i \le l < i+m;\; 0 \le k < n;\;\; a_1(l+k, j+k) \ne 0.$$

Thus $d(i,j)$ is an $m \times n$ rectangular block in A. ☐

COROLLARY 5.2. *Given a matrix A and a positive integer K. The problem of deciding if A has at least K nonoverlapping diagonal blocks is NP-complete.*

COROLLARY 5.3. *Algorithm MNDB-APX is a linear time 2/3-approximation algorithm to find maximum number of nonoverlapping diagonal blocks.*

$$\begin{pmatrix} & x & \\ x & x & x \\ & x & \end{pmatrix} \qquad \begin{pmatrix} x & & x \\ & x & \\ x & & x \end{pmatrix} \qquad \begin{pmatrix} x & x & \\ & x & \\ & x & x \end{pmatrix}$$

(a) (b) (c)

$$\begin{pmatrix} & x & x \\ & x & \\ x & x & \end{pmatrix} \qquad \begin{pmatrix} & & x \\ x & x & x \\ x & & \end{pmatrix} \qquad \begin{pmatrix} x & & \\ x & x & x \\ & & x \end{pmatrix}$$

(d) (e) (f)

FIG. 5.2. *(a) Cross block, (b) diagonal cross block, (c)–(f) jagged cross blocks*

5.2. Cross Blocks. Various regular substructures in a sparse matrix can be exploited to improve memory performance of sparse matrix computations. One possibility is the cross blocks depicted in Fig. 5.2(a). We will identify a cross block with its center, that is we say $c(i,j)$ is a cross block in a matrix A iff A has nonzeros at positions (i,j), $(i,j-1)$,$(i-1,j)$, $(i,j+1)$, and $(i+1,j)$. Below, we prove that finding a maximum number of nonoverlapping cross blocks is NP-complete by using our proof of Theorem 3.4 as a template.

THEOREM 5.4. *Given a matrix A and a positive integer K. The problem of deciding if A has at least K nonoverlapping cross blocks is NP-complete.*

Proof. It is easy to see that this problem can be reduced to the independent set problem, and thus it is in NP. For the NP-completeness proof we use a reduction from the independent set problem on cubic planar graphs. First we use Theorem 2.3 to embed the cubic planar graph onto a grid and then enlarge the grid by 20 as we did for the proof of Theorem 3.4. We can replace each vertex on this grid with a cross pattern in the matrix. Formally, for an $M \times N$ grid, we define a $2M+1 \times 2N+1$ matrix, where grid point (i,j) is replaced by a cross centered at $(2i+1, 2j+1)$ in the matrix. A does not have any other nonzeros besides those in cross blocks corresponding to vertex points. Observe that there are no cross blocks in A, besides those representing grid points. Also observe that unlike the case for rectangular blocks, bends and T-junctions do not cause any problems, since the cross to the left and below the corner vertex of a bend do not overlap.

The only problem is to make sure each edge in G is replaced by an even length path. For this purpose we use the transformation illustrated in Fig. 5.3. Observe that this transformation replaces a chain of odd length with a chain of even length, and consequently making sure of edges in G are replaced with even length paths. $\square$

We can use matrix rotations to reduce the problems of finding other blocks in Fig. 5.2(b–f) to the problem of finding cross blocks as in Fig. 5.2(a). For instance, Rotation 1 transforms jagged crosses, which are illustrated in Fig. 5.2(c) to regular crosses.

THEOREM 5.5. *Given matrix A, let A_1 be its rotated matrix under Rotation 1. $c(i,j)$ is a diagonal cross block in A, iff $c(i+N-j-1, j)$ is a cross block in A_1.*

Proof. The proof only requires applying Rotation 1 to the definition a cross block as for the proof of Theorem 5.1. $\square$

COROLLARY 5.6. *Given a matrix A, and a positive integer K. The problem of*

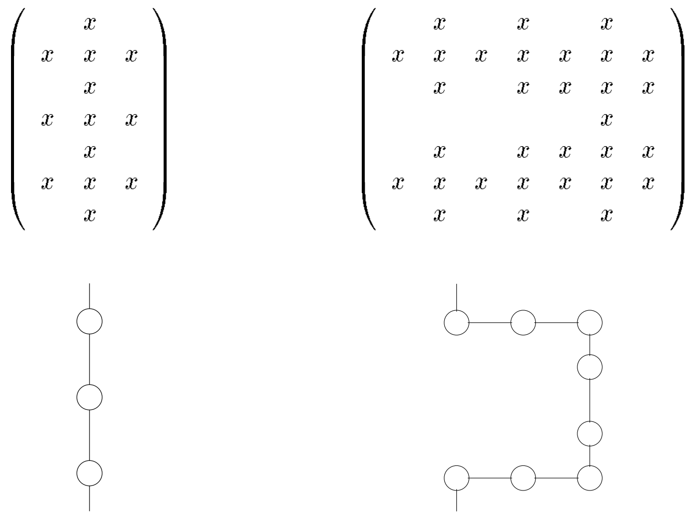

FIG. 5.3. *Odd- to even-length path transformation for cross blocks.*

deciding if A contains at least K nonoverlapping jagged cross blocks is NP-complete.

Similar transformation operators can be transformed for variations of the jagged cross block in Fig. 5.2(d–f). Now we introduce a new rotation operator to transform diagonal cross blocks of Fig. 5.2(b) to regular cross blocks. This rotation is depicted in Fig. 5.1 and below we define it formally.

Rotation 2: *Given an $M \times N$ matrix A, its rotated matrix A_R is an $M+N-1 \times M+N-1$ matrix so that $A(i,j) \neq 0$ iff $A_R(i-j+N-1, i+j)$ for $0 \leq i < M$ and $0 \leq j < N$.*

THEOREM 5.7. *Given matrix A, let A_2 be its rotated matrix under Rotation 2. $c(i,j)$ is a diagonal cross block in A, iff $c(i+N-1-j, i+j)$ is a cross block in A_2.*

Proof. The proof only requires applying Rotation 2 to the definition a cross block. □

COROLLARY 5.8. *Given a matrix A, and a positive integer B. The problem of deciding if A contains at least B nonoverlapping diagonal cross blocks is NP-complete.*

Observe that a greedy algorithm that chooses the leftmost block in the upper most row will yield a 1/2 approximation algorithm for finding cross blocks, and all its variations.

6. Open Problems. This work studies a new problem for the sparse matrix computations community, and brings forth many open problems. One interesting family of problems is the design of heuristics for larger blocks and different substructures, and developing better approximation algorithms. As we discussed in Section 4, our 2/3-approximation algorithm might be generalized for larger blocks, where the runtime complexity will depend on the block size. Another interesting question is if it would be possible to improve the approximation ratio by looking at a larger neighborhood of the leftmost vertex of the uppermost row. Finally, different substructures require different heuristics. For instance the neighborhood structure of the cross block is fairly different than that of the rectangular block, and thus our 2/3-approximation algorithm cannot be applied directly.

Another approach to reduce memory indirections is replacing structural nonzeros of the matrix with numerical zeros. As shown in [10], by selectively replacing structural zeros with numerical zeros, it is possible to gain significant speedups due to better memory performance, even though the number of floating point operations increase. This technique calls for another interesting combinatorial problem. In this case, we need to choose blocks to make sure all nonzeros are covered, and we try to do this by using as few blocks as possible. We call this problem the *minimum block cover problem*, and define it as follows.

> *Given a sparse matrix A, and an oriented substructure α, place minimum number of substructures on A, so that all its nonzeros are covered.*

Notice that this problem is a covering problem, whereas the maximum nonoverlapping substructures problem was an independent set problem. However, the relation between the two problems is not as clear as the relation between the independent set, and vertex cover problems on graphs.

Finally, in this paper we considered finding only one specified structure in the matrix. However it is possible to split a matrix into three or more matrices (e.g., $A = A_d^2 + A_d^1 + A_s$, so that each matrix contains a different substructure. In such a decomposition, the objective will be minimizing the total number of blocks in all matrices. Clearly, this problem is much harder, and even good approximation algorithms (provably or practically) will be valuable.

7. Conclusions.

We studied the problem of finding maximum number of nonoverlapping substructures in a sparse matrix, which we called the *maximum nonoverlapping substructures* problem. Such substructures can be exploited to improve memory performance of sparse matrix operations by reducing the number of memory indirections. We focused on $m \times n$ dense blocks as a substructure (maximum nonoverlapping dense blocks problem) due to their availability in sparse matrices arising in various applications, and effectiveness in decreasing extra load operations. We investigated the relation between the maximum independent set problem and the maximum nonoverlapping substructures problem, and defined a class of graphs where the independent set problem is equivalent to the maximum nonoverlapping dense blocks problem. We used this relation to prove the NP-completeness of the maximum nonoverlapping dense blocks problem. Our proof used a reduction from the maximum independent set problem on cubic planar graphs and adopted orthogonal drawings of planar graphs. We also presented an approximation algorithm for the maximum nonoverlapping dense blocks problem for 2×2 blocks. Our algorithm require only linear time and space, and generate solutions whose sizes are within $2/3$ of the optimal. We also described a $1/2$ approximation algorithm that work for larger blocks and different substructures. We discussed generalizations of our results to different substructures and observed the relation between diagonal blocks and rectangular blocks to show that the two problems are equivalent and one can be reduced to the other by a matrix transformation. We also discussed cross blocks and proved that MNS problem is NP-complete for cross blocks.

REFERENCES

[1] B. Baker, Approximation algorithms for NP-complete problems on planar graphs, *Proc. 24th IEEE Symp. on Foundations of Computer Science* (1983), pages 265–273.

[2] F. Berman, D. Johnson, T. Leighton, P. Shor, L.Snyder, Generalized planar matching, *Journal of Algorithms* **11**, (1990), pages 153–184.

[3] H. de Fraysseix, J. Pach, R. Pollack, How to draw a planar graph on a grid, *Combinatorica*, **10** (1990), pages 41–51.
[4] M. Garey and D. Johnson, *Computers and intractability: A guide to the theory of NP-completeness*, W.H. Freeman and Co., 1979.
[5] E. Im, K. Yelick, R. Vuduc, SPARSITY: An Optimization Framework for Sparse Matrix Kernels, *International Journal of High Performance Computing Applications*, 18 (1), 2004, to appear.
[6] G. Kant, Drawing planar graphs using the canonical ordering, *Algorithmica*, **16**, (1996), pages 4–32.
[7] A. Papakostas and I. Tollis, Algorithms for area-efficient orthogonal drawings. *Computational Geometry* **9** (1998), pages 83–110.
[8] A. Pınar, M. Heath, Improving performance of sparse matrix vector multiplication, *Proc. IEEE/ACM Conf. on Supercomputing 1999*, Portland, OR, November (1999) .
[9] S. Toledo, Improving the memory-system performance of sparse matrix vector multiplication, *IBM Journal of Research and Development*, **41** (6), 1997.
[10] R. Vuduc, J. Demmel, K. Yelick, S. Kamil, R. Nishtala, B. Lee, Performance Optimizations and Bounds for Sparse Matrix-Vector Multiply *Proc. IEEE/ACM Conf. on Supercomputing 2002*, Baltimore, MD, USA, November (2002).